AFHRL-TP-89-61

AIR FORCE

HUMAN RESOURCES

LABORATORY

WORK PERFORMANCE RATINGS: COGNITIVE MODELING AND FEEDBACK PRINCIPLES IN RATER ACCURACY TRAINING

Terry L. Dickinson

Old Dominion University
Department of Psychology
Norfolk, Virginia 23529

Jerry W. Hedge

Universal Energy Systems, Inc.
8961 Tesoro Drive
San Antonio, Texas 78217

Rudolph L. Johnson
Todd A. Silverhart

Old Dominion University
Department of Psychology
Norfolk, Virginia 23529

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

February 1990

Interim Technical Paper for Period June 1985 - September 1989

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

RODGER D. BALLENTINE, Colonel, USAF
Chief, Training Systems Division

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	February 1990	Interim - June 1985 to September 1989

4. TITLE AND SUBTITLE

Work Performance Ratings: Cognitive Modeling and Feedback Principles in Rater Accuracy Training

5. FUNDING NUMBERS

C - F41684-84-D-0020
PE - 62703F
PR - 7719
TA - 18
WU - 40

6. AUTHOR(S)

Terry L. Dickinson
Jerry W. Hedge
Rudolph L. Johnson
Todd A. Silvernhart

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Universal Energy Systems, Inc.
8961 Tesoro Drive, Suite 600
San Antonio, Texas 78217

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Training Systems Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFHRL-TP-89-61

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The present research employed a performance measurement test bed to investigate the influence of training principles on rating accuracy. In one experiment, cognitive modeling principles were utilized to train raters. In comparison to control groups, cognitive modeling training improved rater knowledge of appropriate behaviors to rate, but this training did not improve rating accuracy. In the second experiment, feedback, feedforward, and target score information were manipulated to train raters in the "how and why" of rating. In comparison to controls, this information did not improve rating accuracy. In both experiments, however, rating accuracy was quite high, and it was recommended that the control training procedures should be used in field settings.

14. SUBJECT TERMS

accuracy
cognitive modeling
feedback
feedforward
measurement test bed
rater training
validity
work performance

15. NUMBER OF PAGES

50

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFHRL Technical Paper 89-61 February 1990

WORK PERFORMANCE RATINGS: COGNITIVE MODELING AND FEEDBACK PRINCIPLES IN RATER ACCURACY TRAINING

Terry L. Dickinson

Old Dominion University
Department of Psychology
Norfolk, Virginia 23529

Jerry W. Hedge

Universal Energy Systems, Inc.
8961 Tesoro Drive
San Antonio, Texas 78217

Rudolph L. Johnson
Todd A. Silverhart

Old Dominion University
Department of Psychology
Norfolk, Virginia 23529

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

Martin W. Pellum, Captain, USAF
Chief, Performance Measurement Section
Training Systems Division

Submitted for publication by

Nestor K. Ovalle, Lt Col, USAF
Chief, Training Applications Branch
Training Systems Division

SUMMARY

A role-play exercise from a management skills assessment center was chosen as a medium for investigations of rater accuracy training. The role play required the ratees to deal with a subordinate whose performance was inadequate. In two experiments, raters were trained, and the accuracy of their ratings was evaluated.

The rater training experiments demonstrated that the test bed was especially suited for research on the performance measurement process. All training conditions resulted in accurate ratings. Future research was suggested for varying the information processing demands placed on raters.

PREFACE

This work was conducted in partial fulfillment of Contract No. F41684-84-D-0020 awarded to Universal Energy Systems Incorporated, with the Air Force Human Resources Laboratory (AFHRL). Suzanne Lipscomb served as task monitor. It complements the AFHRL Training Systems Division efforts in job performance criterion development by investigating several innovative approaches to rater training aimed at increasing rater accuracy.

TABLE OF CONTENTS

TABLE OF CONTENTS (concluded)

LIST OF TABLES

WORK PERFORMANCE RATINGS: COGNITIVE MODELING AND FEEDBACK PRINCIPLES IN RATER ACCURACY TRAINING

I. INTRODUCTION

The Armed Forces are engaged in an effort to validate the Armed Services Vocational Aptitude Battery (ASVAB) against on-the-job performance. In order to do this, performance measures must validly and accurately reflect an individual's on-the-job performance. The Air Force's approach is to develop a variety of performance measurement methods and to train appropriate individuals in the use of these methods. The most detailed method, known as Walk-Through Performance Testing, uses work samples to obtain observations of hands-on performance and to test knowledge of task procedures. In addition, task, dimensional, and global rating forms are developed for use by supervisors, peers, and incumbents.

Another aspect of the Air Force's approach is the use of a test bed for investigating methods for training individuals in the use of the measurement methods. The medium chosen for the test bed was a management skills assessment center (Dickinson & Hedge, 1989). In the development of the test bed, particular emphasis was given to dimensions and exercises that had been used frequently in a variety of assessment centers to ensure a domain of performance measures with broad generality to the work setting.

The validity of the test bed's performance measures was established in order to determine the methods that best reflected the dimensions in the performance domain (Dickinson & Hedge, 1989). Eight assessment centers were conducted to generate performance rating measures for an in-basket, two role-play, and two leaderless group discussion exercises. In addition, an experiment was conducted for each type of exercise to determine the construct validity of the dimensions. The results of these three experiments suggested that the conception of a performance dimension be limited to a particular exercise rather than the entire domain of assessment center performance. That is, exercise content and rating format were found to moderate the validity of the dimensions.

An experiment was also conducted to establish target scores for the employee role-play exercise. In this experiment, subject-matter experts were given

enhanced opportunities to rate 10 videotapes of role-play performance. The ratings demonstrated extremely high validity and interrater agreement, indicating that the ratings could be used as target scores for investigations of rating accuracy.

Rater Training Research

Research on rating accuracy has focused on the effects of rater training (e.g., Bernardin & Pence, 1980; Borman, 1977; 1979; Hedge & Kavanagh, 1988; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984). In the training, raters have attended lectures on the nature of distortions in ratings, participated in small group discussions of distortions in ratings, and learned the performance behaviors associated with target ratings. The training is often presumed to result in more accurate performance measurement (Zedeck & Cascio, 1984). However, as Spool (1978) points out, most rater training research fails to consider major learning principles in the design of the training program such as modeling, practice, feedback, and transfer of training.

The purpose of the present research was to investigate the influence of learning principles on rating accuracy. The performance test bed was utilized in two experiments to assess the influence of modeling and feedback training on rating accuracy.

II. EXPERIMENT 1: COGNITIVE MODELING IN RATER TRAINING

Although behavior modeling has been advocated as an effective approach for training supervisory skills (Decker & Nathan, 1985; Goldstein & Sorcher, 1974; Kraut, 1976), it cannot be directly applied to improve rating accuracy. The covert nature of performance rating requires the verbal presentation of the cognitive strategies of the expert rater (i.e., the model).

The research evidence for cognitive modeling has been limited primarily to clinical applications such as reducing test anxiety (e.g., Bruch, 1978; Meichenbaum, 1972). However, McIntyre and Bentson (1984) have investigated the influence of cognitive modeling training on observation accuracy. In their research, an expert observer described the behaviors relevant to effective teaching while a videotape of a college lecturer was being shown. In subsequent viewing of

videotapes, cognitive modeling training was found to increase the proportion of behaviors correctly reported by observers. Unfortunately, their research did not investigate rating accuracy.

A model must be perceived by the raters as an expert on the rating process. Clearly, a model can be shown to be expert by the knowledge demonstrated in descriptions of appropriate ratings. This knowledge is communicated in the performance rating context by "thinking aloud" (McIntyre & Bentson, 1984). However, expertise may be further enhanced by having experience in the performance context. For example, in the employee role play (Dickinson & Hedge, 1989), the expert rater could also be depicted as having been a role player.

This experiment compared the influence of observer and role-player cognitive modeling strategies on the accuracy of performance ratings.

Method

Raters

The raters were 27 male and 25 female, undergraduate business students attending Old Dominion University. Raters were paid $40.00 for their participation. They ranged in age from 19 to 33, and their mean age was 22.3 years.

Design

The design included four training conditions (i.e., no training, dimension and behavioral checklist training, observer cognitive modeling, and role-player cognitive modeling). The dimension and behavioral checklist training gave raters in-depth familiarity with the rating formats that the experts used to develop the target scores (see Dickinson & Hedge, 1989). Along with the no-training condition, it served as a control condition to evaluate the effects of cognitive modeling.

Each condition was administered to a group of 12 to 15 raters. A group rated the videotaped performance of 9 ratees on the employee role play for the dimensions of problem analysis, problem solution, and sensitivity.

The experiment was conducted in two sessions. During session 1, the training conditions were

administered to the raters. One of the 10 videotaped performances was used to practice the rating procedure. This session lasted from 2 1/2 to 4 1/2 hours, depending on the training condition. Raters returned the next day and rated the remaining 9 videotaped performances. Session 2 lasted 2 1/2 hours.

Procedure

For all conditions, training was accomplished with a videotape of instructions and demonstrations. The videotape was interrupted at various points (a) to answer rater questions or (b) to allow the raters to study materials that were distributed.

For all conditions, a basic procedure was followed. Raters were told that they would be viewing videotapes of the role-play performances of potential managers, who had been participants in a management assessment center. Following a brief description of management assessment centers, the scenario of the employee role play was distributed to the raters and discussed by the trainer. Next, the behaviorally anchored rating scale (BARS) and behavioral checklist formats were distributed, and the trainer described how to use these formats for rating the videotaped performances. Further, an enactment of the role play was shown to all raters for training. In this training role play and the 10 videotaped role-play performances, the trainer played the role of the ratee. The training role play was also used to administer the cognitive modeling conditions, but for the remaining conditions, it was used only to familiarize raters with the role play. Finally, one of the 10 videotaped role plays was viewed and rated for practice by all raters. Checklist ratings were made as the videotape was viewed, and raters were instructed to review these ratings prior to using the BARS. Specific variations of the basic procedure are described in the following paragraphs.

The no-training condition required 2 1/2 hours to administer. Raters were given time to become familiar with the checklist and the BARS prior to being instructed on the proper use of the formats. In the remaining conditions, raters were given dimension and behavioral checklist training prior to being instructed on the proper use of the BARS format.

In the instructions on use of the BARS format, raters were informed that the five behavioral statements

for each dimension were meant to represent five different levels of performance. Raters were told to circle the statement that best reflected the level of performance on each dimension. Instructions were also provided to clarify the "could be expected" format of the BARS. It was emphasized that raters should select the statement which described the level of performance that was consistently demonstrated by the ratee in the videotape. Raters were told that although some of the behaviors to be observed in the videotapes were statements on the BARS, this did not indicate that these statements should necessarily be selected as a rating. Rather, they were told to consider all behaviors relevant to a dimension in the videotape before making their ratings.

In dimension and behavioral checklist training, the dimensions were first defined and discussed by the trainer. Next, each of the 45 behavioral checklist items was read, and where appropriate, a description was provided as to how the items should be interpreted in relation to the role play. Finally, the BARS were distributed, and each scale anchor on the BARS was read. Further, instructions were given on the proper use of the BARS format. This training required 3 1/2 hours.

Observer and role-player cognitive modeling were administered after the viewing of the training role-play videotape. The script for the role play was carefully prepared to include 40 of the 45 possible behavioral items on the checklist. The trainer discussed each of the 45 checklist behaviors and verbalized a rationale for its appearance or absence. This discussion of the behaviors was enhanced by replaying segments of the videotape to illustrate relevant behaviors. Finally, raters were also instructed on the proper use of the BARS format. Modeling training required 4 1/2 hours.

For the observer cognitive modeling condition, the trainer modeled the expert rater from the perspective of an observer of the performance. For each dimension, the trainer gave an evaluation of the ratee's handling of the interview in terms of each checklist item and each BARS statement. For each dimension, the trainer also announced a BARS rating and provided a summary rationale for that rating. In sum, the trainer described a rationale for dimension ratings by "thinking aloud" (McIntyre & Bentson, 1984).

For the role-player cognitive modeling condition, the trainer modeled the expert rater from the perspective of having played the role of the subordinate. The trainer discussed the consistent perspective that the trainer had taken as the subordinate in the role plays. For example, when asked a question about a problem area by the manager (i.e., the ratee), the subordinate would respond in a manner to suggest that the subordinate was unaware of the problem area. For each dimension, the trainer gave an evaluation of the ratee's handling of the interview in terms of (a) each checklist item and (b) each BARS statement. For each dimension, the trainer also announced a BARS rating and provided a summary rationale for that rating.

Questionnaires

Participants completed pre-training and post-training questionnaires during session 1 to evaluate the efficacy of training. For these questionnaires, participants were instructed to match behavioral statements to their illustrative dimension (i.e., problem analysis, problem solution, or sensitivity).

A third questionnaire was administered at the beginning of session 2. This pre-rating questionnaire was used to: (a) determine if training remained effective and (b) refamiliarize participants with dimensions and behavioral statements prior to rating. Copies of the matching questionnaires are contained in Appendix A.

A questionnaire was also administered at the conclusion of session 2. This questionnaire was used to assess perceptions of rating accuracy and expertise of the trainer. A copy of the post-rating questionnaire is contained in Appendix B.

Results

Training Checks

Analyses were conducted on the pre-training, post-training, and pre-rating questionnaires to assess the ability of raters to match behavior statements correctly to illustrative dimensions. Each analysis was based on the one-way design for the factor of training conditions. A priori contrasts were formed to compare the training conditions. The means for the training conditions and the F-ratios for the contrasts are shown in Table 1.

The pre-training analysis indicated that prior to training the raters did not differ significantly ($p > .05$) in their ability to match statements to dimensions. Following training, however, the raters in the training conditions improved in their ability to match statements compared to those in the no-training condition. This improvement was demonstrated in the first session immediately following training ($p < .01$) and in the second session prior to viewing videotapes ($p < .05$).

Basic Accuracy

An analysis of variance procedure was used to evaluate the accuracy of the ratings (Dickinson, 1987). In addition, variance components and intraclass correlation coefficients (Vaughn & Corballis, 1969) were computed to compare the amounts of rating variance accounted for by the sources of variation.

The design included the factors from the basic accuracy design (i.e., rating sources, dimensions, and ratees) as repeated measures. For each rater, orthonormal contrasts were formed between the ratings and corresponding target scores (i.e., the rating sources). These 27 contrasts described variation due to discrepancies between ratings and target scores for the 9 ratees for each of the 3 dimensions. The design also included the factor of training conditions. Although counterintuitive, accuracy of ratings was demonstrated by a lack of statistical significance and, of course, small discrepancies between the ratings and target scores. A summary of the results of the analysis is included in Table 2.

The results indicated inaccuracies in the ratings for the factors from the basic accuracy design. The significant effect for Rating Sources reflected that raters tended to rate ($M = 3.01$) higher than warranted by the target scores ($M = 2.87$). More importantly, inaccuracies in the ratings occurred for dimensions, ratees, and their interaction.

The Dimensions effect only accounted for 2% of the rating variance. Tukey's honestly significant difference (HSD) procedure revealed that the mean discrepancies between ratings and target scores for problem analysis ($M = .24$) and problem solution ($M = .20$) were significantly greater than the mean discrepancy for sensitivity ($M = -.13$).

Table 1. Means and F-Ratios for Pre-Training, Post-Training, and Pre-Rating Contrasts Between Training Conditions

	Training conditions				Contrasts[a]		
Questionnaire	OM	RM	DT	NT	C1	C2	C3
Pre-Training	19.2	18.8	18.6	17.8	1.08	.17	.07
Post-Training	20.4	19.7	20.5	17.8	10.43**	1.42	1.37
Pre-Rating	20.8	21.0	20.2	18.8	4.92*	.76	.03

Note. OM, observer cognitive modeling; RM, role-player cognitive modeling; DT, dimension and behavioral checklist training; and NT, no training. C1, contrast between no-training and remaining conditions; C2, contrast between dimension and behavioral checklist training condition and cognitive modeling training conditions; and C3, contrast between role-player and observer cognitive modeling conditions.

[a]Degrees of freedom for F-ratios were 1 and 48.

*p <.05. **p <.01.

T-tests were also performed on the mean discrepancies for each of the dimensions to detect significance from zero. Each t-test was evaluated against a p-level of p <.0014. This conservative p-level maintained a family error rate of p <.05 for the set of t-tests conducted for the basic accuracy effects of Dimensions, Ratees, and Dimensions by Ratees.

The mean discrepancies for problem analysis and problem solution were significantly different from zero, while the discrepancy for sensitivity was not.

The Ratees effect was also significant, and it accounted for 4% of the rating variance. Tukey's procedure revealed that the mean discrepancy for ratee 3 (M = .34) was significantly greater than the discrepancies for ratees 7 (M = .03), 1 (M = .02), 2 (M = -.09), and 6 (M = -.19). It also revealed that the discrepancies for ratees 4 (M = .27), and 8 (M = .26)

Table 2. Analysis of Variance for Training Conditions on the Accuracy of Ratings

Source	df	MS	F-ratio	VC	ICC
Rating Sources (S)	1	14.920	10.32*	.028	.04
Train Conds (C)	3	.857	.41[a]	-.003	.00
Raters/C (R/C)	45	1.445	2.41*	.031	.04
Dimensions (D)	2	18.430	2.67[a]	.018	.02
D x C	6	1.360	3.62[a]	.004	.01
D x R/C	90	.419	1.07	.006	.01
Ratees (E)	8	4.700	7.85*	.028	.04
E x C	24	1.223	2.04*	.017	.02
E x R/C	360	.599[b]			
D x E	16	6.883	17.65*	.135	.19
D x E x C	48	.347	.89	-.004	.00
D x E x R/C	720	.390[b]			

Note. For experiments 1 and 2, if a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the source's coefficient was set to zero. Train Conds, training conditions; VC, variance component; ICC, intraclass correlation coefficient.

[a]Quasi F-ratio.

[b]Pooled to estimate a residual variance component equal to .460 for computing intraclass correlation coefficients.

*$p < .01$.

were significantly greater than those for ratees 6 and 2. Further, the discrepancies for ratees 3, 4, and 8 differed significantly from zero.

The Dimensions by Ratees interaction accounted for the largest amount of rating variance (i.e., 19%). The mean discrepancies for the ratees on each of the dimensions are shown in Table 3. Tukey's procedure indicated that for problem analysis the mean discrepancy (a) for ratee 3 was significantly greater than those for ratees 6, 8, and 4, (b) for ratee 2 was significantly greater than those of ratees 6 and 8, and (c) for ratee 6

Table 3. T-tests for Mean Discrepancies of Zero Between Ratings and Target Scores for the Dimensions by Ratees Interaction

	Problem analysis		Problem solution		Sensitivity	
Ratee	MD	t-test	MD	t-test	MD	t-test
1	.24	2.12	.25	2.83	-.43	4.05*
2	.61	4.65*	-.32	3.62*	-.56	5.14*
3	.74	5.79*	.23	1.97	.05	.72
4	.06	.68	.70	6.66*	.06	.74
5	.50	3.74*	-.24	2.69	.03	.34
6	-.56	5.35*	.13	2.27	-.13	1.29
7	.18	1.78	-.17	1.66	.09	.83
8	.01	.15	.56	6.29*	.20	2.25
9	.39	3.10	.69	7.43*	-.49	5.58*

Note. MD, mean discrepancy between ratings and target scores. T-tests were based on 48 degrees of freedom.

*p <.0014.

significantly less than those of all ratees. For problem solution, the mean discrepancies for ratees 4, 9, and 8 were significantly greater than the discrepancies for ratees 2, 5, and 7. For sensitivity, the mean discrepancy (a) for ratee 2 was significantly less than the discrepancies for ratees 8, 7, 4, 3, and 5, (b) for ratee 9 was significantly less than the discrepancies for ratees 8 and 7, and (c) for ratee 1 was less than that for ratee 8.

In general, the significant mean discrepancies between ratees varied widely by dimension. Ratees 2 and 9 were noteworthy in this regard. On problem analysis the discrepancies were positive, and on sensitivity the discrepancies were negative. However, on problem solution ratee 2 had a negative difference, and ratee 9 had a positive difference.

Finally, t-tests for the significance of the mean discrepancies from zero also reflected the complexity of the Ratees by Dimensions interaction. As shown in Table 3, both problem analysis and problem solution had 4 discrepancies that differed significantly from zero, while sensitivity had 3. Although the dimensions had a similar number of mean discrepancies different from zero, only ratees 2 and 9 appeared to be poorly rated on all dimensions.

Training Conditions

The interactions of the basic accuracy factors with the training conditions were of particular concern in this experiment. These interactions reflected the ability of training to moderate rating inaccuracies.

As shown in Table 2, only the Ratees by Training Conditions interaction was statistically significant ($p < .01$). The mean discrepancies for the ratees for each of the training conditions are shown in Table 4. Tukey's procedure indicated that in observer cognitive modeling the mean discrepancy for ratee 8 was significantly greater than the discrepancies of ratees 1, 6, and 2. In the role-player cognitive modeling and dimension and behavioral checklist training conditions, no significant discrepancies occurred. In the no-training condition, the mean discrepancies for ratees 3 and 5 were significantly greater than those of ratees 6, 7, 8, and 2.

T-tests for significance of mean discrepancies from zero were also conducted. Each t-test was evaluated against a p-level of $p < .0014$. This p-level maintained a family error rate of $p < .05$ for the set of t-tests conducted for the Ratees by Training Conditions interaction. The t-tests are shown in Table 4. In general, the raters were able to rate quite accurately within each of the training conditions. Only the mean discrepancies for ratee 4 within the role-player cognitive modeling and dimension and behavioral checklist training conditions were statistically discrepant from zero. In these instances, raters tended to rate higher than warranted by the target scores.

Post-Rating Questionnaire

Analyses were conducted on the post-rating questionnaire items to assess raters' perceptions of

Table 4. T-tests for Mean Discrepancies of Zero Between Ratings and Target Scores for the Ratees by Training Conditions Interaction

	Observer cognitive modeling		Role-player cognitive modeling		Dimension & behavior training		No training	
Ratee	MD	t-test	MD	t-test	MD	t-test	MD	t-test
1	-.21	1.70	.14	.63	-.08	.65	.28	1.94
2	-.12	.84	-.12	.89	-.09	.64	-.03	.25
3	.29	2.26	.14	.79	.28	2.10	.61	3.74
4	.25	2.37	.42	5.89*	.27	4.25*	.17	1.43
5	-.09	1.03	-.05	.69	-.05	.32	.59	2.54
6	-.20	1.82	-.14	1.03	-.22	1.73	-.18	1.43
7	-.01	.07	.21	1.16	.09	.80	-.15	1.18
8	.44	3.66	.47	3.27	.25	2.42	-.09	.56
9	.12	.76	.14	.72	.24	1.66	.26	2.06

Note. MD, mean discrepancy between ratings and target scores. T-tests for the conditions were based on 11 degrees of freedom for observer cognitive modeling, 9 for role-player cognitive modeling, 14 for dimension and behavioral checklist training, and 11 for no training.

*p <.0014.

rating accuracy and the trainer. The item means and F-ratios for a priori contrasts among the means are shown in Table 5.

The results suggested that raters perceived observer cognitive modeling to help rating accuracy significantly more than the remaining conditions. That is, the contrast between the no-training condition and the remaining conditions was significant as well as the contrast between observer and role-player cognitive modeling conditions (p <.05). A post hoc comparison between the observer cognitive modeling and dimension and behavioral checklist training conditions approached significance ($F = 3.51$; $df = 1, 48$; p <.07).

Table 5. Means and F-Ratios for Post-Rating Contrasts Between Training Conditions

Questionnaire item	Training conditions				Contrasts[a]		
	OM	RM	DT	NT	C1	C2	C3
Training help accuracy?	4.1	3.4	3.5	3.1	4.02*	.93	4.22*
Knowledgeable trainer?	4.5	3.8	3.9	3.7	2.18	1.28	3.87
Ratings are accurate?	3.4	3.2	3.4	3.1	1.32	.11	.44

Note. OM, observer cognitive modeling; RM, role-player cognitive modeling; DT, dimension and behavioral checklist training; and NT, no training. C1, contrast between no-training and remaining conditions; C2, contrast between dimension and behavioral checklist training condition and cognitive modeling training conditions; and C3, contrast between role player and observer cognitive modeling conditions.

[a]Degrees of freedom for F-ratios were 1 and 48.

*$p < .05$.

Raters also tended to see the trainer as more knowledgeable in the observer cognitive modeling condition. The a priori contrast between observer and role-player cognitive conditions approached significance ($p < .06$). Support for the interpretation was also provided by post hoc comparisons of the observer cognitive modeling condition to the no-training condition ($F = 6.04$; $df = 1, 48$; $p < .05$) and to the dimension and behavioral checklist training condition ($F = 3.88$; $df = 1, 48$; $p < .06$).

Discussion

The training conditions did not differ appreciably in the accuracy of the performance ratings. These results do not agree with those obtained in previous rater training research (e.g., McIntyre & Bentson,

1984; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1984). Comparisons with previous research offer interpretations for the present results.

The amount of time allocated to training in the present experiment differed substantially from the amount in previous research. The no-training condition required the least amount of time to administer (i.e., 2 1/2 hours). In contrast, the most extensive training in previous rater accuracy research (Pulakos, 1984, 1986) required less time to administer (i.e., 1 1/2 hours) than the present control condition. Further, the no-training conditions administered in previous research (McIntyre, et al. 1984; Pulakos, 1984, 1986) consisted only of 5 minutes of general instructions given to familiarize raters with rating procedures. In contrast, raters in the present no-training condition were given (a) time to study the checklist and BARS formats, (b) the role-play scenario and shown a training videotape to familiarize them with the rating context, (c) instructions on the proper use of the formats, and (d) opportunity to practice rating one videotape. Apparently, the no-training condition in the present experiment gave all raters sufficient time to become as proficient in making accurate ratings as raters in the training conditions.

The relevant research on cognitive modeling suggested that raters could be trained to use the cognitive strategy of the expert rater. The "thinking aloud" technique was used successfully by McIntyre and Bentson (1984) to improve observational accuracy. Of course, observation precedes the integration and decision making that are subsequently required to make performance ratings, suggesting that the "thinking aloud" technique may not be sufficient to improve rating accuracy.

The research on behavior modeling is more extensive (Decker & Nathan, 1985), and the components for successful behavior modeling include: (a) demonstration of appropriate behaviors by an expert, (b) practice of these behaviors, (c) feedback regarding reproduction of appropriate behaviors, and (d) transfer of training through continued repetition of appropriate behaviors. The covert nature of making performance ratings requires an adaptation of these components. The "thinking aloud" procedure and subsequent practice in rating of the training videotape were adaptations of the first three

components. However, in retrospect, there was no assurance that raters reproduced the appropriate cognitive strategies. Raters were not required to reproduce their cognitive strategies "publicly." Future research should investigate the effects on rating accuracy of a public rehearsal of cognitive strategy by the rater.

Although few differences were obtained in rating accuracy between the training conditions, mean discrepancies between ratings and target scores were generally small in magnitude. For example, only 2 discrepancies were greater than .50 for the 9 ratees in the 4 training conditions (see Table 4). Nonetheless, inaccuracies did occur. Most troublesome were the mean discrepancies for the ratees within each of the dimensions. The ratings for ratees 2 and 9 were noteworthy in the nature of their discrepancies from the target scores. Apparently, raters could rate overall performance quite accurately, but they had more difficulty rating dimension performance accurately. Future research should investigate explanations and training solutions.

The observer cognitive modeling approach was superior in influencing rater perceptions. This training approach was perceived by raters to help them evaluate accurately, and it led to perceptions of greater trainer expertise. In contrast, the role-player cognitive modeling was not perceived to be a superior approach. Perhaps, "thinking aloud" from the role player's point of view lost salience because the trainer was observed to be the role player on the videotapes in all conditions.

Finally, the implications of this experiment are that a choice among the present training conditions for use in field settings should be based on a practical considerations. In field settings where raters are familiar with the rating context, only training on the use of the rating formats needs to be given. This training should include an opportunity for practice ratings such as occurred in the no-training condition.

In the next experiment, raters were given the opportunity to make practice ratings, and information about the target scores was varied to determine its influence on rating accuracy.

III. EXPERIMENT 2: INFORMATION TYPE AND MODE IN RATER TRAINING

The usefulness and importance of feedback for improving performance is well recognized in laboratory (e.g., Adams, 1968; Ammons, 1956) and organizational settings (Ilgen, Fisher, & Taylor, 1979; Sassenrath, 1975). Historically, feedback has been considered information received by an individual about past behavior that provides an indication of the accuracy or correctness of a response (Annett, 1969). However, Ilgen et al. (1979) have noted that feedback also has value as information concerning the "how and why" that underlies performance. In rater training research, feedback has contained "how and why" information (Dickinson & Silverhart, 1986; McIntyre et al., 1984; Pulakos, 1984, 1986). Target scores, themselves, contain information about the accuracy of ratings. However, feedback can also inform raters of a behavioral rationale for the target scores. One purpose of this experiment was to compare the effectiveness of target score and behavioral rationale information as well as their combination.

In complex cognitive tasks, information regarding the outcome of a given decision may be less optimal for improving performance than information regarding the structure for making a decision (Adelman, 1981; Hammond, McClelland, & Mumpower, 1980). It has been suggested that providing information about task structure before decision making may be an effective strategy for learning complex cognitive tasks (Bogart, 1980; Hendrix & Dudycha, 1981). This technique is referred to as feedforward (Bjorkman, 1972). This experiment also compared the influence of feedback and feedforward information in rater training on rating accuracy.

Method

Raters

The raters were 47 male and 49 female, graduate and undergraduate business students attending Old Dominion University. Raters were paid $40.00 for their participation. They ranged in age from 18 to 41, and their mean age was 21.7 years.

Design

The design included type and mode of information conditions as well as two control conditions. Three types of information (i.e., target score, behavioral rationale, and combination of target score and behavioral rationale) were crossed with two modes of information presentation (i.e., feedback and feedforward). In addition, dimension training and no-training conditions were included in the design.

Each of the research conditions was administered to a group of 12 raters. After this, each group rated the reenacted videotaped performances of 7 ratees on the employee role play for the dimensions of problem analysis, problem solution, and sensitivity.

Three of the 10 reenacted videotapes were used to administer the information conditions. All conditions are described in the following paragraphs.

Information Type Conditions. Raters in the target score information conditions were shown the mean expert ratings (i.e., target scores) on an overhead projector for each videotape of role-play performance. Raters in the behavioral rationale information conditions were provided a lecture describing the relevant behaviors that were taken into consideration by the expert raters in determining their ratings. This lecture was based on the checklist of behaviors agreed to by the consensus of the experts. In addition, a videotape consisting of segments that illustrated the relevant behaviors for each role-play performance was shown to the raters. The combination target score and behavioral rationale conditions consisted of a presentation of all information contained in the target score and behavioral rationale conditions.

Information Mode Conditions. Feedback information was presented after the raters viewed and rated each of the three videotaped performances. In the feedforward conditions, all information was provided before viewing each of the three videotapes.

Control Conditions. The dimension training condition consisted of training on dimension definitions, proper use of the BARS, and familiarization with the behavioral anchors on the BARS. This "basic" training was also included in the information training

conditions. The no-training condition consisted of training only on the proper use of the BARS. However, raters were given time to read and become familiar with dimension definitions and behavioral anchors. In the two control conditions, no information was provided regarding the role-play performances displayed in the three videotapes. However, the raters did make practice ratings of the videotapes.

The combinations of information type and mode conditions required 4 1/2 hours to administer, while the no-training control and dimension training conditions required 2 1/2 and 3 1/2 hours, respectively.

Procedure

For all conditions, training was accomplished with a videotape of the instructions and demonstrations. As in Experiment 1, raters were oriented to the purpose of the research. Following the description of management assessment centers, the videotape was interrupted, and a copy of the employee role play was distributed to the raters and discussed by the trainer. Next, the videotape was restarted and an enactment of the employee role play was shown. The trainer played the role of the ratee.

After the enactment, the dimensions for rating performance were described, except for the no-training condition in which raters were given time to become familiar with the dimensions. The description included a discussion of the dimension definitions and examples of behaviors relevant to each dimension. These example behaviors were selected from the checklist.

Next, instructions on the proper use of the BARS format were given. These instructions were identical to those administered in Experiment 1.

Following the BARS instructions, the manipulations of information type and mode occurred. A condition-specific procedure was employed to ensure that raters attended to the information. In the target score feedback condition, raters were told to plot their ratings and the target scores on a graph after viewing a videotape. In the target score feedforward condition, raters were told simply to plot the target scores before viewing a videotape. In the behavioral rationale feedback condition, after the behavioral rationale was

provided, raters were told to complete the checklist so they could indicate which behaviors they considered in making ratings. In the behavioral rationale feedforward condition, raters were told to complete the checklist as they viewed the videotape. In the combination target score and behavioral rationale conditions, the plotting and checklist procedures for feedback and feedforward were combined.

Next, the three videotapes were viewed and rated for practice. While viewing each videotape, raters took notes on behaviors believed to be relevant to problem analysis, problem solution, and sensitivity. After viewing a videotape, raters used the BARS to rate performance on the dimensions. This completed the first session, and it lasted from 2 1/2 to 4 1/2 hours, depending on research condition.

Participants returned the following day for a second session. Refresher training was provided on the dimension definitions and use of the rating scales. Further, raters were reminded to take notes as they viewed the 7 experimental videotapes. The second session lasted 2 hours.

Questionnaires

Participants completed pre-training and post-training questionnaires during session 1 to evaluate the efficacy of training. Further, a pre-rating questionnaire was administered at the beginning of session 2. The questionnaires were identical to those used in Experiment 1.

A questionnaire was also administered at the conclusion of session 2. This questionnaire was used to assess reactions to the training. A copy of the post-rating questionnaire is contained in Appendix C.

Results

Training Checks

Analyses of variance were conducted on the pre-training, post-training, and pre-rating questionnaires to assess the number of behavior statements correctly matched by raters to illustrative dimensions. Each analysis was based on a one-way design. Three a priori contrasts were formed to compare the conditions.

First, the no-training condition was compared to the information training conditions. Second, the dimension training condition was compared to the information training conditions. A final contrast compared the no-training and dimension training conditions. The means for the conditions and the F-ratios for the contrasts are shown in Table 6.

The pre-training analysis indicated that prior to training, the raters did not differ significantly ($p > .05$) in the number of statements correctly matched to dimensions. Immediately following training, the raters in the information training conditions did improve in their ability to match statements to the dimensions compared to the no-training condition ($p < .05$). However, in the second session prior to viewing videotapes, raters in the conditions were similar in their ability to match statements correctly.

Basic Accuracy

An analysis of variance procedure was used to evaluate the accuracy of the ratings, and it included the factors from the basic accuracy design as repeated measures. For each rater, orthonormal contrasts were formed to describe variation due to discrepancies between ratings and target scores for the 7 ratees for each of the 3 dimensions. The design also included factors of information type, information mode, and contrasts for the control groups. The no-training (NT) contrast compared the no-training condition to the remaining conditions, and the dimension training (DT) contrast compared the dimension training condition to the information conditions. A summary of the results of the analysis is included in Table 7.

The results indicated inaccuracies in the ratings for the factors from the basic accuracy design. The significant effect for Rating Sources reflected that raters tended to rate ($M = 3.23$) higher than warranted by the target scores ($M = 2.81$). More importantly, inaccuracies in the ratings occurred for dimensions, ratees, and their interaction.

The Dimensions effect accounted for 13% of the rating variance. Tukey's HSD procedure revealed that the mean discrepancies between ratings and target scores were significantly different for all dimensions.

Table 6. Means and F-Ratios for Pre-Training, Post-Training, and Pre-Rating Contrasts Between Research Conditions

	Training conditions			Contrasts[a]		
Questionnaire	MI	DT	NT	C1	C2	C3
Pre-Training	18.4	17.9	17.6	.98	.37	.08
Post-Training	19.7	19.1	18.2	7.61*	1.15	1.66
Pre-Rating	21.4	19.9	20.4	.00	.44	.00

Note. MI, mean of information training conditions; DT, dimension training condition; NT, no-training condition. C1, contrast between no-training and information training conditions; C2, contrast between dimension training condition and information training conditions; and C3, contrast between no-training and dimension training conditions.

[a]Degrees of freedom for F-ratios were 1 and 88.

*$p < .01$.

T-tests were also performed on the mean discrepancies for each of the dimensions to detect significance from zero. Each t-test was evaluated against a p-level of $p < .0016$ to maintain a family error rate of $p < .05$ for the basic accuracy effects.

The mean discrepancy for problem analysis ($M = .85$) and for problem solution ($M = .68$) were significantly different from zero, while the discrepancy for sensitivity ($M = -.26$) was not different from zero.

The Ratees effect was also significant, accounting for 13% of the rating variance. Tukey's procedure revealed that (a) the mean discrepancy for ratee 1 was significantly greater than the discrepancies for the remaining ratees, and (b) except for ratees 1 and 6, the mean discrepancy for ratee 3 was significantly greater than those for the remaining ratees. Further, only the discrepancies for ratee 1 ($M = 1.42$) and ratee 3 ($M = .62$) differed significantly from zero.

Table 7. Analysis of Variance for Information Types and Modes on the Accuracy of Ratings

Source	df	MS	F-ratio	VC	ICC
Rating Sources (S)	1	180.459	137.34*	.089	.08
Info Types (T)	2	1.282	1.50[a]	.001	.00
Info Modes (M)	1	1.112	1.32[a]	.000	.00
T x M	2	1.234	.77[a]	-.000	.00
Dim Training (DT)	1	.037	.02[a]	-.002	.00
No Training (NT)	1	2.375	1.56[a]	.000	.00
Raters/Cond (R/C)	88	1.314	1.82	.028	.02
Dimensions (D)	2	120.088	8.39*[a]	.140	.13
D x T	4	.278	.37[a]	-.001	.00
D x M	2	.308	.64[a]	-.000	.00
D x T x M	4	.106	.19[a]	-.001	.00
D x DT	2	2.809	6.08*[a]	.004	.00
D x NT	2	.090	.24[a]	-.000	.00
D x R/C	176	.538	1.57	.028	.02
Ratees (E)	6	32.254	44.67*	.146	.13
E x T	12	.260	.36	-.006	.00
E x M	6	.251	.35	-.004	.00
E x T x M	12	1.008	1.40	.008	.01
E x DT	6	1.036	1.43	.005	.00
E x NT	6	.932	1.29	.003	.00
E x R/C	528	.722[b]			
D x E	12	14.119	41.16*	.191	.17
D x E x T	24	.553	1.61	.009	.01
D x E x M	12	.286	.83	-.002	.00
D x E x T x M	24	.353	1.03	.001	.00
D x E x DT	12	.271	.79	-.004	.00
D x E x NT	12	.175	.51	-.007	.00
D x E x R/C	1056	.343[b]			

Note. Info, information; Dim, dimensions; Cond, conditions; VC, variance component; ICC, intraclass correlation coefficient.

[a]Quasi F-ratio.

[b]Pooled to estimate a residual variance component equal to .469 for computing intraclass correlation coefficients.

*$p < .01$.

The Dimensions by Ratees interaction accounted for the largest amount (i.e., 17%) of rating variance. Tukey's procedure indicated that for problem analysis, ratees 1 and 3 accounted for 11 of 15 significant differences in mean discrepancies and for problem solution for 9 of 15. In contrast, for sensitivity ratees 2 and 7 accounted for all of the 11 significant differences.

Finally, t-tests for the significance of the mean discrepancies from zero also reflected the influence of the sensitivity dimension on the nature of the interaction. As shown in Table 8, 6 of 7 mean discrepancies between the rating and target scores for ratees differed significantly from zero for problem analysis and problem solution, while only 2 of 7 differed significantly for sensitivity.

Research Conditions

The interactions of the basic accuracy factors with the research conditions (i.e., information type, information mode, and control conditions) reflected the ability of the research conditions to moderate rating inaccuracies. As shown in Table 7, only the interaction with Dimensions of the contrast comparing the mean discrepancies of the dimension training condition to those of the six information training conditions was statistically significant ($p < .01$). The main and interaction effects of information type and mode as well as of the contrast comparing the no-training condition to the remaining research conditions were not significant.

Further, a contrast comparing the dimension training and no-training conditions and its interactions with information mode and type were computed in a post hoc analysis. Only the interaction of this contrast with Dimensions was statistically significant ($F = 4.17$; $df = 2, 31$; $p < .05$).

Inspection of the mean discrepancies between ratings and target scores for the dimension training condition compared to the information training conditions revealed a linear by linear interaction. Scheffe's post hoc procedure indicated that the interaction effect was due to a greater linear slope for the information training conditions ($p < .01$). That is, greater mean discrepancies occurred for the information

Table 8. T-tests for Mean Discrepancies of Zero Between Ratings and Target Scores for the Dimensions by Ratees Interaction

	Problem analysis		Problem solution		Sensitivity	
Ratee	MD	t-test	MD	t-test	MD	t-test
1	1.69	29.80*	1.22	16.60*	.09	1.59
2	.42	5.15*	.32	4.10*	-.51	6.62*
3	.92	9.64*	.30	4.07*	.10	1.78
4	-.03	.39	.43	6.26*	.01	.20
5	.40	5.87*	-.16	2.20	.01	.10
6	.21	3.60*	.62	9.59*	-.14	1.74
7	.58	6.84*	.64	8.51*	-.85	10.54*

Note. MD, mean discrepancy between ratings and target scores. T-tests were based on 95 degrees of freedom.

*p <.0016.

training conditions compared to the dimension training condition on problem analysis (M = .85 versus .68), problem solution (M = .69 versus .50), and sensitivity (M = -.32 versus -.01). Similarly, the no-training condition had a greater linear slope than the dimension training condition (p <.01), as reflected by its greater dimension means (i.e., M = 1.02, .82, and -.18, respectively).

Post-Rating Questionnaire

A summary of the analyses of the post-rating questionnaire items is shown in Table 9. Raters did not differ in their perceptions concerning the helpfulness of the research conditions to rate accurately nor the accuracy of their ratings (p >.05). However, they did perceive that the information provided in the information conditions to be more understandable compared to the no-training condition (p <.05).

Table 9. Means and F-Ratios for Post-Rating Contrasts Between Research Conditions

Questionnaire item	Training conditions			Contrasts[a]		
	MI	DT	NT	C1	C2	C3
Training help accuracy?	4.25	4.17	4.08	.71	.18	.10
Training understandable?	4.40	4.08	3.92	5.31*	2.30	.36
Ratings are accurate?	3.83	4.00	3.67	.78	.78	1.81

Note. MI, mean of information training conditions; DT, dimension training condition; NT, no-training condition. C1, contrast between no-training and information training conditions; C2, contrast between dimension training condition and information training conditions; and C3, contrast between no-training and dimension training conditions.

[a]Degrees of freedom for F-ratios were 1 and 88.

*$p < .05$.

Discussion

The mode and type of information presented in rater training did not influence the accuracy of the performance ratings. These results are not consisent with those obtained in previous research on rater training, and similar to Experiment 1, the nature and amount of time allocated to training may explain the failure to establish differences between the research conditions.

In Experiment 1, however, raters in the training conditions compared to the no-training condition were able to match statements to dimensions more accurately on the post-training and pre-rating questionnaires. In the present experiment, raters in the research conditions matched statements more accurately only on

the post-training questionnaire. Perhaps the brief refresher training provided at the beginning of session 2 and prior to administration of the pre-rating questionnaire was sufficient additional training for raters in the no-training condition to improve their matching of statements.

In Experiment 2 (and other rater training research), the cues available to raters were behavioral in nature and occurred in conjunction with other irrelevant behaviors. In previous research on cognitive tasks (e.g., Adelman, 1981; Lindell, 1976; Nystedt & Magnusson, 1973), the cues tended to be numerical in nature and relevant to decision making. In addition, feedforward information was more quantitative in previous research, consisting of statistical information on the weighting of cues in order to make accurate judgments. That is, the decision maker had to evaluate the numerical cues and arrive at a prediction of the target score. The simplest performance rating task involves observation in addition to evaluation (Thornton & Zorich, 1980), and in more complicated rating tasks, additional cognitive processes such as encoding or recall are involved (Feldman, 1981). Perhaps feedforward information for training the raters did not improve accuracy due to the additional cognitive demands of the rating task. Relevant and irrelevant behavioral cues needed to be processed, and these cues were not in a simple numerical format.

A comparison of the rating accuracy obtained in the present experiment to that obtained in Experiment 1 indicates that the training procedures utilized in Experiment 1 lead to greater rating accuracy. For example, a comparison of the intraclass correlation coefficients (see Tables 2 and 7) reveals that raters were more accurate in Experiment 1 for Rating Sources (ICC = .04 versus .08), Dimensions (ICC = .02 versus .13), and Ratees (ICC = .04 versus .13). Further, the mean discrepancies for the Ratees by Dimensions interactions (see Tables 3 and 8) also reflect the greater accuracy obtained in Experiment 1.

A potential explanation for the greater rater accuracy may be the behavioral checklist training that was administered in Experiment 1 in the dimension and behavioral checklist and no-training conditions. The control conditions in Experiment 2 did not receive checklist training, because this training would have

confounded these control conditions with the information type and mode conditions. That is, the behavioral checklist was used to manipulate information type and mode. Perhaps, the behavioral checklist is an important heuristic in rater training.

IV. CONCLUSIONS

The two rater training experiments indicated that cognitive modeling and information on the "how and why" of performance do not influence rating accuracy. Although previous research suggested that they could be important principles in rater training, the no-training conditions yielded ratings that were similar in rating accuracy. Although the no-training conditions provided much more thorough and extensive training than the no-training conditions utilized in previous research investigations, this is not a sufficient explanation for the failure of cognitive modeling and performance information. Future research should address the components needed to utilize cognitive modeling and performance information principles for effective rater training.

At present, it is recommended that training in operational settings employ the no-training condition administered in Experiment 1. The training currently used by the Air Force in obtaining supervisor, peer, and incumbent ratings would appear to satisfy these requirements. Current Air Force training includes (a) familiarization with rating forms and distortions in ratings, and (b) opportunity for and feedback on practice ratings, with the majority of allotted time concentrated on procedures included in the Experiment 1, no-training condition.

REFERENCES

Adams, J.A. (1968). Response feedback and learning. Psychological Bulletin, 70, 486-504.

Adelman, L. (1981). The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. Organizational Behavior and Human Performance, 27, 423-442.

Ammons, R.B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. The Journal of General Psychology, 54, 279-299.

Annett, J. (1969). Feedback and human behavior. Baltimore, MD: Penguin Books.

Bernardin, H.J., & Pence, E.C. (1980). The effects of rater training: Creating new response sets and decreasing accuraacy. Journal of Applied Psychology, 65, 60-66.

Bjorkman, M. (1972). Feedforward and feedback as determiners of knowledge and policy: Notes on a neglected issue. Scandinavian Journal of Psychology, 13, 152-158.

Bogart, D.H. (1980). Feedback, feedforward, and feedwithin: Strategic information in systems. Behavioral Science, 25, 237-249.

Borman, W.C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 258-272.

Borman, W.C. (1979). Format and training effects on rating accuracy, and rater errors. Journal of Applied Psychology, 64, 410-421.

Bruch, M.A. (1978). Type of cognitive modeling, imitation of modeled tactics, and modification of test anxiety. Cognitive Therapy and Research, 2, 147-164.

Decker, P.J., & Nathan, B.R. (1985). Behavior modeling training: Principles and applications. New York: Praeger Publishers.

Dickinson, T.L. (1987). Designs for evaluating the validity and accuracy of performance ratings. Organizational Behavior and Human Decision Processes, 40, 1-21.

Dickinson, T.L., & Hedge, J.W. (1989). Work performance ratings: Measurement test bed for validity and accuracy research (AFHRL-TP-88-36, AD-A205 165). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Dickinson, T.L., & Silverhart, T.A. (1986, August). Training to improve the accuracy and validity of performance ratings. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Goldstein, A.P., & Sorcher, M. (1974). Changing supervisor behavior. New York: Pergamon Press.

Hammond, K.R., McClelland, G.H., & Mumpower, J. (1980). Human judgment and decision making: Theories, methods, & procedures. London: Praeger.

Hedge, J.W., & Kavanagh, M.J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. Journal of Applied Psychology, 73, 68-73.

Hendrix, W.H., & Dudycha, A.L. (1981). Feedforward and feedback in multiple cue probability learning -- facilitating or debilitating? Journal of Experimental Education, 49, 137-146.

Ilgen, D.R., Fisher, C.D., & Taylor, M.S. (1979). Consequences of individual feedback on behavior in organizations. Journal of Applied Psychology, 64, 349-371.

Kraut, A.I. (1976). Developing managerial skills via modeling techniques: Some positive research findings -- a symposium. Personnel Psychology, 29, 325-328.

Lindell, M.K. (1976). Cognitive and outcome feedback in multiple-cue probability learning tasks. Journal of Experimental Psychology: Human Learning and Memory, 2, 739-745.

McIntyre, R.M., & Bentson, C.A. (1984, August). A comparison of methods for training behavioral observation: Modeling works! Paper presented at the annual meeting of the American Psychological Association, Toronto.

McIntyre, R.M., Smith, D.E., & Hassett, C.E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Meichenbaum, D.H. (1972). Cognitive modification of test-anxious college students. Journal of Consulting and Clinical Psychology, 39, 370-380.

Nystedt, L., & Magnusson, D. (1973). Cue relevance and feedback in a clinical prediction task. Organizational Behavior and Human Performance, 9, 100-109.

Pulakos, E.D. (1984). A comparison of training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Pulakos, E.D. (1986). The development of training programs to increase accuracy with different rating tasks. Organizational Behavior and Human Decision Processes, 38, 76-91.

Sassenrath, J.M. (1975). Theory and results on feedback and retention. Journal of Educational Psychology, 67, 894-899.

Spool, M.D. (1978). Training programs for observers of behavior: A review. Personnel Psychology, 31, 853-888.

Thornton, G.C., III, & Zorich, S. (1980). Training to improve observer accuracy. Journal of Applied Psychology, 65, 351-354.

Vaughn, G.M., & Corballis, M.J. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. Psychological Bulletin, 72, 204-213.

Zedeck, S., & Cascio, W.F. (1984). Psychological issues in personnel decisions. Annual Review of Psychology, 35, 461-518.

APPENDIX A: PRE-TRAINING, POST-TRAINING, AND PRE-RATING QUESTIONNAIRES

I. Pre-Training Questionnaire

Before you begin training, we would like to gather some preliminary information. In collecting this information, you will become familiar with the dimensions and the behaviors involved in the research. Your responses will not be used to evaluate your individual performance in this research. It is simply one way we can establish the effectiveness of training. The questions should take approximately 10 minutes to complete. We ask that you give careful consideration to your responses. Please answer all questions.

You are asked to match each behavioral item with a performance dimension. For each behavioral item, choose the performance dimension that you think best represents that behavior and write the letter of that dimension in the space preceding the behavior.

Performance Dimensions

A. Problem Analysis B. Problem Solution C. Sensitivity

Behavioral Items
(The dimension for an item is indicated in parentheses.)

Outlines a plan of action that the employee should have followed. (Problem solution)

Relates the employee's adjustment to the new store to problems that the employee is experiencing. (Problem analysis)

Inquires whether the employee had ever received any complaints from subordinates but goes no further with this information. (Problem analysis)

Compliments the employee on feelings of job responsibility. (Sensitivity)

Acknowledges that a lot of employees are apprehensive about the appraisal process. (Sensitivity)

Asks the employee's opinion of what could be done to improve the employee's relations with subordinates. (Problem analysis)

Recommends that the employee exert more authority and let the staffers know who is boss. (Problem solution)

Asks whether the employee told subordinates about the employee's standards for quality work. (Problem analysis)

Inquires whether the employee checked last year's inventory before ordering the picnic tables. (Problem analysis)

Tells the employee that he or she wants to make the employee's performance even better. (Sensitivity)

Inquires whether the employee has had any problems adjusting to the store. (Problem analysis)

States that he/she has confidence in the employee. (Sensitivity)

Inquires whether the employee has any questions about job responsibilities. (Problem analysis)

Suggests that the employee could threaten to reduce the hours of staffers if they did not do their jobs. (Problem solution)

Suggests that the employee show subordinates what to do rather than the employee doing it. (Problem solution)

Recommends that the employee try delegating more responsibility to subordinates without explaining how. (Problem solution)

Expresses the desire to work with the employee to remedy the problems. (Sensitivity)

Inquires whether the employee's subordinates needed more training. (Problem analysis)

Inquires whether the employee has any problems with subordinates. (Problem analysis)

Inquires as to the reason the employee works so many hours but does not use the response to the question to address a problem. (Problem analysis)

Outlines action plans for employee development. (Problem solution)

Suggests that the employee needs to take time to do a better job with scheduling and ordering. (Problem solution)

Listens intently to what the employee has to say. (Sensitivity)

II. Post-Training Questionnaire

We have completed the training component of this research. We are now interested in determining how effective this training has been in enabling you to distinguish between performance dimensions. Therefore, we would like you to complete this questionnaire before you return to session 2 for the rating task. Once again, your answers will not be used to evaluate your performance in this study. It is simply a means by which we can establish what you have learned from this training experience. The questions should take approximately 10 minutes to complete. We ask that you give careful consideration to your responses. Please answer all questions.

You are asked to match each behavioral item we have discussed with a performance dimension. For each behavioral item, choose the performance dimension that you think best represents that behavior and write the letter of that dimension in the space preceding the behavior.

Performance Dimensions

A. Problem Analysis B. Problem Solution C. Sensitivity

Behavioral Items

(The dimension for an item is indicated in parentheses.)

Suggests that the employee sit down with subordinates and attempt to develop a better working relationship. (Problem solution)

Inquires whether the employee consulted subordinates regarding their scheduling preferences. (Problem analysis)

Acknowledges that the employee's past performance appraisals were good. (Sensitivity)

Suggests that the employee explain to the staffers how the inventory system works. (Problem solution)

Acknowledges that it is difficult to turn over responsibility. (Sensitivity)

Says that the employee is ultimately responsible for ensuring that all of the work is done properly. (Sensitivity)

Acknowledges the difficulty of adjusting to a larger store. (Sensitivity)

Asks the employee about thoughts and feelings of the issues that had been discussed. (Sensitivity)

Puts the employee at ease by asking how the employee likes being at the new store. (Sensitivity)

Investigates how the employee took care of the problem when subordinates didn't do the work or didn't do it well. (Problem analysis)

Doesn't thank the employee at the conclusion of the interview. (Sensitivity)

Suggests that the employee hand out note cards with responsibilities listed on them to subordinates to deal with the delegation problem. (Problem solution)

Suggests the employee talk with subordinates and find out how they feel about working nights and weekends. (Problem solution)

Inquires about the reason that the employee believes subordinates are not doing their work. (Problem analysis)

Suggests that a goal could be obtained without specifying the manner in which it could be accomplished. (Problem solution)

Suggests that if the staffers did not want to work nights and weekends that the employee should rotate them. (Problem solution)

Conveys the impression that the employee is guilty until proven innocent. (Sensitivity)

Relates the employee's lack of patience in dealing with subordinates to the employee's long hours. (Problem analysis)

Indicates being impressed by all of the hours the employee has been working. (Sensitivity)

Suggests that the employee might want to share knowledge so that subordinates had a better understanding of how the company works. (Problem solution)

Suggests that the employee has to develop better communications with subordinates without explaining how. (Problem solution)

Begins the interview by asking if there is anything that the employee would like to bring up, and then doesn't use the information to initiate a line of questioning for some problem. (Problem analysis)

III. Pre-Rating Questionnaire

Before you begin the rating task, we would like to assess again the effectiveness of training and to re-acquaint you with the dimensions and behaviors. As in the two previous questionnaires, your answers will not be used to evaluate your individual performance in this research. The questions should take approximately 10 minutes to complete. We ask that you give careful consideration to your responses. Please answer all questions.

You are asked to match each behavioral item with a performance dimension. For each behavioral item, choose the performance dimension that you think best represents that behavior and write the letter of that dimension in the space preceding the behavior.

Performance Dimensions

A. Problem Analysis B. Problem Solution C. Sensitivity

Behavioral Items
(The dimension for an item is indicated in parentheses.)

Inquires whether the employee has any questions about job responsibilities. (Problem analysis)

Says that the employee is ultimately responsible for ensuring that all of the work is done properly. (Sensitivity)

Relates the employee's lack of patience in dealing with subordinates to the employee's long hours. (Problem analysis)

Listens intently to what the employee has to say. (Sensitivity)

Suggests that if the staffers did not want to work nights and weekends that the employee should rotate them. (Problem solution)

Suggests that the employee talk with subordinates and find out how they feel about working nights and weekends. (Problem solution)

Acknowledges the difficulty of adjusting to a larger store. (Sensitivity)

Suggests that the employee hand out note cards with responsibilities listed on them to subordinates to deal with the delegation problem. (Problem solution)

Acknowledges that a lot of employees are apprehensive about the appraisal process. (Sensitivity)

Asks the employee whether subordinates were told about work standards in response to the employee's comments about the poor quality of subordinates' work. (Problem analysis)

Inquires whether the employee had ever received any complaints from subordinates but goes no further with this information. (Problem analysis)

Investigates how the employee took care of the problem when subordinates didn't do the work or didn't do it well. (Problem analysis)

Tells the employee that he or she wants to make the employee's performance even better. (Sensitivity)

Recommends that the employee try delegating more responsibility to subordinates without explaining how. (Problem solution)

Recommends that the employee might want to share knowledge so that subordinates have a better understanding of how the company works. (Problem solution)

Expresses the desire to work with the employee to remedy the problems. (Sensitivity)

Doesn't thank the employee at the conclusion of the interview. (Sensitivity)

Inquires as to the reason the employee works so many hours but does not use the response to the question to address a problem. (Problem analysis)

Inquires whether the employee has had any problems adjusting to the store. (Problem analysis)

Suggests that the employee could threaten to reduce the hours of the staffers if they did not do their jobs. (Problem solution)

Relates the employee's adjustment to the new store to the problems being experienced. (Problem analysis)

Inquires whether the employee checked last year's inventory before ordering the picnic tables. (Problem analysis)

Suggests that the employee explain to the staffers how the inventory system works. (Problem solution)

APPENDIX B: POST-RATING QUESTIONNAIRE FOR EXPERIMENT 1

Please circle the response alternative that reflects your reaction to the this research project.

1. To what extent did the training help you to evaluate the ratee accurately?

Not at all	Somewhat	Quite a bit	To a great extent	Completely
1	2	3	4	5

2. To what extent did you perceive the trainer as knowledgeable in observation and performance rating?

Not at all	Somewhat	Quite a bit	To a great extent	Completely
1	2	3	4	5

3. To what extent was the experiment a learning experience for you?

Not at all	Somewhat	Quite a bit	To a great extent	Completely
1	2	3	4	5

4. How confident are you that your ratings are accurate measures of the individual's performance?

Not at all	Somewhat	Quite a bit	To a great extent	Completely
1	2	3	4	5

APPENDIX C: POST-RATING QUESTIONNAIRE FOR EXPERIMENT 2

Please circle the response alternative that reflects your reaction to the this research project.

1. To what extent did the information presented in Session 1 help you to evaluate the ratee accurately?

Not at all	Somewhat	Quite a bit	To a great extent	Completely
1	2	3	4	5

2. To what extent was the information presented in Session 1 understandable?

Not at all	Somewhat	Quite a bit	To a great extent	Completely
1	2	3	4	5

3. To what extent was the experiment a learning experience for you?

Not at all	Somewhat	Quite a bit	To a great extent	Completely
1	2	3	4	5

4. To what extent do you feel confident that your ratings are accurate measures of the assessees' performance?

Not at all	Somewhat	Quite a bit	To a great extent	Completely
1	2	3	4	5